Gene Discovery and Functional Analysis of Human Genetic Variation in Disease-Related Transcription Pathways

Author(s): **Douglas A. Bell, Xuting Wang, Han-Yo Cho, Brian Chorley, Steve Kleeberger**

Affiliation(s): **National Institute of Environmental Health Sciences**

Background

Human genetic variation, such as single nucleotide polymorphism (SNPs), can play an important role in determining susceptibility to environmental stress. In particular, polymorphisms occurring in transcription factor binding sites may change the binding of transcription factors and modulate gene expression in an allele-specific manner. The NRF2 protein binds to a sequence called the Antoxidant Response Element (ARE) in the regulatory regions of oxidative stress responsive genes.

Goal: Identify human polymorphisms that alter NRF2 binding.

The system is described in Figure 2. It relies on NCBI dbSNP, gene, and genome databases, and utilizes gene expression datasets from our collaborators. A set of PERL and SQL programs have been implemented to:

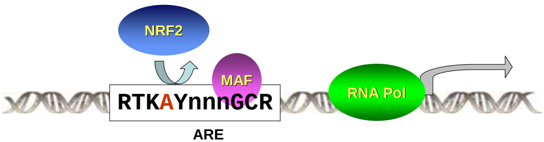
1. Construct a position weight matrix (PWM) model for searching novel AREs in the human genome.
2. Identify SNPs whose sequences fit the ARE motif.
3. Map the SNPs to regulatory regions of human gene.
4. Examine the evolutionary conservation of by phylogenetic footprinting.
5. Select the oxidant stress inducible genes by mining microarray expression profiles.
6. Analyze association between genotypes of ARE SNPs and expression phenotypes of target genes.
7. Test SNPs in disease association studies.

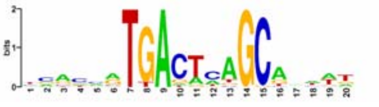
Figure 1

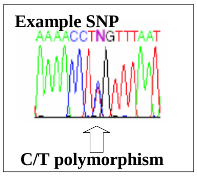
NRF2 mediates transcriptional activation of target genes by binding to an Antioxidant Response Element (ARE) sequence in upstream promoter region.

We are identifying human polymorphisms that alter transcription factor binding and regulation of gene expression.

Code: R=A/G; W= A/T; K= G/T; Y= C/T; S= G/C



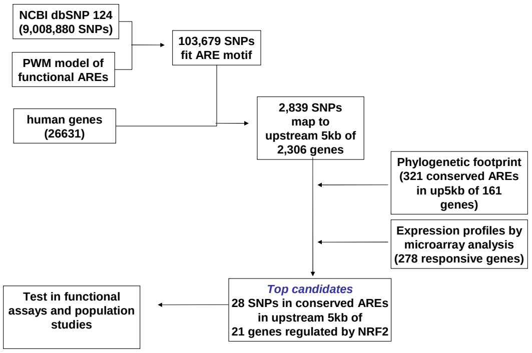



Example SNP

C/T polymorphism

Figure 2

Figure 2. Identification of ARE SNPs. The chart shows the procedure and intermediate results when our integrated discovery system is applied to detect ARE SNPs from 9 millions of uniquely mapped SNPs in human genome. Our top candidates are shown in Table 1.


NCBI dbSNP 124 (9,008,880 SNPs)

PWM model of functional AREs

103,679 SNPs fit ARE motif

human genes (26631)

2,839 SNPs map to upstream 5kb of 2,306 genes

Phylogenetic footprint (321 conserved AREs in up5kb of 161 genes)

Expression profiles by microarray analysis (278 responsive genes)

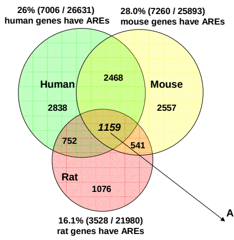
Top candidates
28 SNPs in conserved AREs in upstream 5kb of 21 genes regulated by NRF2

Test in functional assays and population studies

Figure 3

Response elements that are conserved across multiple species are more likely to be functional. Phylogenetic footprinting is used to identify conservation.

Survey AREs in upstream 5kb regions



(UPPER) Survey AREs in upstream 5kb regions of all genes in human, mouse, and rat genomes using position weight matrix, and found 1159 orthologous genes have AREs in upstream.

(LOWER) Identification of conserved AREs by examining the multiple sequence alignment of upstream 5kb regions of orthologous genes.

Among 1159 ARE genes, 161 genes have 321 conserved AREs.
Multiple sequence alignment is from UCSC genome center.

Align upstream 5kb of orthologous genes

Example: Thioredoxin gene

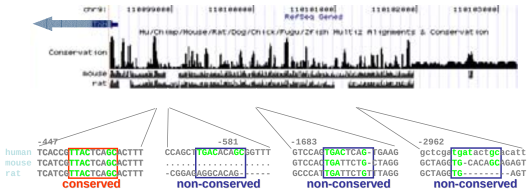

chr9:
RefSeq Genes
Hu/Chimp/Mouse/Rat/Dog/Chick/Fugu/Zfish Multiz Alignments & Conservation
Conservation
mouse
rat
-447
-581
-1683
-2962
human TCACCGTTACTCAGCACTTT CCAGCTTGACACAGCGGTTT GTCCAGTGACTCAG-TGAAG gctcgatgatactgcacatt
mouse TCATCGTTACTCAGCACTTT .... GTCCACTGATTCTG-CTAGG GCTAGGTG-CACAGCAGAGT
rat TCATCGTTACTCAGCACTTT -CGGAGAGGCACAG------ GCCCATTGATTCTGTTTAGG GCTAGGTG---------AGT
conserved
non-conserved
non-conserved
non-conserved

Table 1. Results

Table 1 shows ten candidates sorted on ΔPWM. The red letters are SNPs, and the green letters are core nucleotides of ARE motif. A SNP in a core position causes larger ΔPWM, and greater predicted impact.

The genes associated with the SNPs below are implicated in the *in vivo* antioxidant mechanism.

These candidates are currently being evaluated in functional assays and population studies.

SNP sequence RTKAYnnnGCR	Offset	Gene ontology	Allele 1 max PWM	Allele2 min PWM	ΔPWM
ccactgWgactttgcccattg	-4522	xenobiotic metabolism	11.65	7.02	4.62
TGCTTGMGACTAAGCCAGACC	-2357	electron transport	9.51	4.88	4.62
aaaaaaNgactcagaatgaca	-821	glutathione transferase	9.88	5.26	4.62
GGCTTCTGACTCAYTGAAATA	-4918	oxidative stress	8.51	3.88	4.62
tctctttgaatctgYcacttt	-200	xenobiotic metabolism	8.19	3.61	4.58
cagacatcactaagYctcagt	-1927	oxidative stress	8.02	3.43	4.58
AGGGCTTGARTATGCTTCCTG	-2892	hydrolase activity	6.82	3.79	3.03
AGGCTCTGASTCTGCTTCCGC	-1085	acute-phase response	7.76	5.39	2.37
GAAACGTGACTYGGGGCTATA	-1059	glutathione metabolism	9.06	7.34	1.72
ggaggctgaatcagcatgSga	-3146	oxidoreductase activity	9.00	8.34	0.65

SNP is Red; Core is in Green; R=A/G; W= A/T; K= G/T; Y= C/T; S= G/C

NIEHS
National Institute of Environmental Health Sciences

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY

epascienceforum 2006
Your Health • Your Environment • Your Future